SYSTEM AND METHOD FOR PERFORMING EFFICIENT DOCUMENT SCORING AND CLUSTERING

Field of the Invention

The present invention relates in general to concept and term scoring and clustering and, in particular, to a system and method for performing efficient document scoring and clustering.

Background of the Invention

Large collections of documents have become increasingly available in electronically stored form due, in part, to the widespread adoption of computer-automated information and decision support systems. At the same time, electronically stored document collections have increasingly complemented and often supplanted traditional forms of printed communications. Electronically stored documents present several significant advantages over traditional printed formats, including efficient storage, rapid searchability, and facilitating immediate communication and publication over networking means, including the Internet.

From a pragmatic standpoint, the availability of electronically stored document collections has presented both a treasure and a curse to those seeking information discovery and retrieval. These types of document collections have expanded to include various forms of information classes, such as word processing documents, electronic mail, Worldwide Web (or simply "Web") pages, spreadsheets, databases, and the like. And although now available in a highly searchable format, information embedded in documents stored in an electronic format must generally still be "mined" at a semantic level to discover and retrieve the data contained within. Mining out the semantic content of a document collection is essential to certain fields of endeavor, such as during the discovery phase of litigation. However, efficiently discovering and extracting such

embedded semantic information can be an intractable problem, particularly when the size of the collection of documents is large.

Text mining is at the core of the information discovery process, and is described in D. Sullivan, "Document Warehousing and Text Mining, Techniques for Improving Business Operations, Marketing, and Sales," Chs. 1-3, Wiley Computer Publishing (2001), the disclosure of which is incorporated by reference. Text mining involves the compiling, organizing and analyzing of document collections to support identification of types of information contained in the documents and to discover relationships between relevant facts. However, identifying relevant information can be difficult. First, extracting relevant content requires a high degree of precision and recall. Precision is the measure of how well the documents returned in response to a query actually address the query criteria. Recall is the measure of what should have been returned by the query. Typically, the broader and less structured the documents, the lower the degree of precision and recall. Second, analyzing an unstructured document collection without the benefit of *a priori* knowledge in the form of keywords and indices can present a potentially intractable problem space. Finally, synonymy and polysemy can cloud and confuse extracted content. Synonymy refers to multiple words having the same meaning and polysemy refers to a single word with multiple meanings. Fine-grained text mining must reconcile synonymy and polysemy to yield meaningful results.

Text mining is a significant first step in the overall process of discovering semantic meanings within a document collection. A further problem involves classifying the documents within a collection with respect to *ad hoc* categories of interest. For instance, during the discovery phase of litigation, documents must often be categorized into distinct groups, such as "relevant," "non-relevant," and "privileged." Generally, the various documents falling into each group share certain characteristics, which can often be expressed as concepts and terms.

Similarly, categorizing the documents themselves into groups of related documents may be necessary as an aid to post-text mining document analysis. Text mining creates a multi-dimensional problem space that can be difficult to

intuitively comprehend based on the presence of concepts and terms within the document collection overlapping by various degrees. Data visualization tools are available to display groups or "clusters" of documents, such as described in commonly-assigned U.S. patent applications, Serial No. 09/944,475, filed August 31, 2001, pending, and Serial No. 09/943,918, filed August 31, 2001, pending, and Serial No. 10/084,401, filed February 25, 2002, pending, the disclosures of which are incorporated by reference. Data visualization tools enable a user to rapidly comprehend and pare down the potential search field within a document collection, based on extracted concepts and terms.

In the prior art, text mining is performed in two ways. First, syntactic searching provides a brute force approach to analyzing and extracting content based on literal textual attributes found in each document. Syntactic searching includes keyword and proximate keyword searching as well as rule-based searching through Boolean relationships. Syntactic searching relies on predefined indices of keywords and stop words to locate relevant information. However, there are several ways to express any given concept. Accordingly, syntactic searching can fail to yield satisfactory results due to incomplete indices and poorly structured search criteria.

A more advanced prior art approach uses a vector space model to search for underlying meanings in a document collection. The vector space model employs a geometric representation of documents using word vectors. Individual keywords are mapped into vectors in multi-dimensional space along axes representative of query search terms. Significant terms are assigned a relative weight and semantic content is extracted based on threshold filters. Although substantially overcoming the shortcomings of syntactic searching, the multivariant and multidimensional nature of the vector space model can lead to a computationally intractable problem space. As well, the vector space model fails to resolve the problems of synonymy and polysemy.

Therefore, there is a need for an approach to identifying semantic information within a document collection based on extracted concepts and terms.

Preferably, such an approach would assign a score to each concept and term based on the inherent characteristics of each document and the overall document set.

There is a further need for an approach to clustering documents within a document collection with respect to similarities reflected by the scores assigned to the concepts and terms. Preferably, such an approach would accept a set of candidate seed documents for evaluation and initial clustering.

Summary of the Invention

The present invention provides a system and method for scoring and clustering documents based on extracted concepts and terms. Canonical concepts are formed from concepts and terms extracted from a set of documents and the frequencies of occurrences and reference counts of the concepts and terms are determined. Each concept and term is then scored based on frequency, concept weight, structural weight, and corpus weight. The scores are compressed and assigned to normalized score vectors for each of the documents. A similarity between each normalized score vector is determined, preferably as a cosine value. A set of candidate seed documents is evaluated to select a set of seed documents as initial cluster centers based on relative similarity between the assigned normalized score vectors for each of the candidate seed documents. The remaining non-seed documents are evaluated against the cluster centers also based on relative similarity and are grouped into clusters based on a best fit, subject to a minimum fit criterion.

An embodiment provides a system and method for grouping clusters of semantically scored documents. A score is determined and assigned to at least one concept extracted from a plurality of documents based on at least one of a frequency of occurrence of the at least one concept within at least one such document, a concept weight, a structural weight, and a corpus weight. Clusters of the documents are formed by applying the score for the at least one concept to a best fit criterion for each such document.

A further embodiment provides a system and method for providing efficient document scoring of concepts within a document set. A frequency of occurrence of at least one concept within a document retrieved from the document

set is determined. A concept weight is analyzed reflecting a specificity of meaning for the at least one concept within the document. A structural weight is analyzed reflecting a degree of significance based on structural location within the document for the at least one concept. A corpus weight is analyzed inversely weighing a reference count of occurrences for the at least one concept within the document. A score associated with the at least one concept is evaluated as a function of the frequency, concept weight, structural weight, and corpus weight.

Still other embodiments of the present invention will become readily apparent to those skilled in the art from the following detailed description, wherein are described embodiments of the invention by way of illustrating the best mode contemplated for carrying out the invention. As will be realized, the invention is capable of other and different embodiments and its several details are capable of modifications in various obvious respects, all without departing from the spirit and the scope of the present invention. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not as restrictive.

Brief Description of the Drawings

FIGURE 1 is a block diagram showing a system for performing efficient document scoring and clustering, in accordance with the present invention.

FIGURE 2 is a block diagram showing the system modules implementing the document analyzer of FIGURE 1.

FIGURE 3 is a data flow diagram showing the stages of document scoring performed by the document analyzer of FIGURE 1.

FIGURE 4 is a data flow diagram showing the stages of document clustering performed by the document analyzer of FIGURE 1.

FIGURE 5 is a flow diagram showing a method for performing efficient document scoring and clustering, in accordance with the present invention.

FIGURE 6 is a flow diagram showing the routine for performing document parsing for use in the method of FIGURE 5.

FIGURE 7 is a data structure diagram showing a schema for a document record maintained in the database of FIGURE 1.

FIGURE 8 is a data structure diagram showing a schema for a concept record maintained in the database of FIGURE 1.

FIGURE 9 is a data structure diagram showing a schema for an associated concept record maintained in the database of FIGURE 1.

FIGURE 10 is a data structure diagram showing a schema for a content record maintained in the database of FIGURE 1.

FIGURE 11 is a flow diagram showing a routine for comparing documents for use in the method of FIGURE 5.

FIGURE 12 is a flow diagram showing a routine for scoring concepts and terms for use in the routine of FIGURE 11.

FIGURE 13 is a graph showing, by way of example, the frequency of concept references.

FIGURE 14 is a flow diagram showing a routine for forming clusters for use in the method of FIGURE 5.

FIGURE 15 is a flow diagram showing a routine for applying a dynamic threshold for use in the routine of FIGURE 14.

FIGURE 16 is a graph diagram showing, by way of example, a dynamic threshold in a cluster of documents.

Detailed Description

Glossary

Keyword: A literal search term, which is either present or absent from a document. Keywords are not used in the evaluation of documents as described herein.

Term: A normalized root stem of a single word appearing in the body of at least one phrase.

Phrase: Two or more words co-occurring in the body of a document.

Concept: A collection of terms or phrases defining a specific meaning.

Theme: Two or more concepts defining a semantic meaning.

Cluster: Documents identified to contain a common theme.

The foregoing terms are used throughout this document and, unless indicated otherwise, are assigned the meanings presented above.

System Overview

FIGURE 1 is a block diagram showing a system 10 for performing efficient document scoring and clustering, in accordance with the present invention. By way of illustration, the system 10 operates in a distributed computing environment, which includes a plurality of heterogeneous systems and document sources. The system 10 includes a production server 11, which executes a workbench application 15 for providing a framework for acquiring, logging, culling, and preparing documents for automated review and analysis. The workbench application 15 includes a document analyzer 31 for performing efficient document scoring and clustering, as further described below with reference to FIGURE 2. The production system 11 is coupled to a storage device 13, which stores documents 14, in the form of structured or unstructured data, and a database 30 for maintaining document information.

The document analyzer 31 analyzes documents retrieved from a plurality of local sources. The local sources include documents 17 maintained in a storage device 16 coupled to a local server 15 and documents 20 maintained in a storage device 19 coupled to a local client 18. The local server 15 and local client 18 are interconnected to the production system 11 over an intranetwork 21. In addition, the document analyzer 31 can identify and retrieve documents from remote sources over an internetwork 22, including the Internet, through a gateway 23 interfaced to the intranetwork 21. The remote sources include documents 26 maintained in a storage device 25 coupled to a remote server 24 and documents 29 maintained in a storage device 28 coupled to a remote client 27.

The individual documents 17, 20, 26, 29 include all forms and types of structured and unstructured data, including electronic message stores, such as word processing documents, electronic mail (email) folders, Web pages, and graphical or multimedia data. Notwithstanding, the documents could be in the form of organized data, such as stored in a spreadsheet or database.

In the described embodiment, the individual documents 17, 20, 26, 29 include electronic message folders, such as maintained by the Outlook and Outlook Express products, licensed by Microsoft Corporation, Redmond, Washington. The database is an SQL-based relational database, such as the Oracle database management system, release 8, licensed by Oracle Corporation, Redwood Shores, California.

The individual computer systems, including production system 11, server 15, client 18, remote server 24 and remote client 27, are general purpose, programmed digital computing devices consisting of a central processing unit (CPU), random access memory (RAM), non-volatile secondary storage, such as a hard drive or CD ROM drive, network interfaces, and peripheral devices, including user interfacing means, such as a keyboard and display. Program code, including software programs, and data are loaded into the RAM for execution and processing by the CPU and results are generated for display, output, transmittal, or storage.

Document Analyzer

FIGURE 2 is a block diagram showing the system modules 40 implementing the document analyzer 31 of FIGURE 1. The document analyzer 31 includes four modules: parsing 41, scoring 42, clustering 43, and display and visualization 44. The parsing module 41 processes documents 14 retrieved from the storage device 13 into document records 48, concept records 49, term records 50, and content records 51, which are maintained in the database 30, as further described below with reference to FIGURE 6. The parsing module 41 optionally utilizes a global stop concept (GSC) cache 45 to selectively filter out global concepts.

The scoring module 42 generates scores 52 for each of the concepts and terms, based on frequencies 53, concept weights 54, structural weights 55, and corpus weights 56, as further described below with reference to FIGURE 11. Briefly, the frequencies 53 indicate the number of occurrences of a given concept or term within a document 14. The concept weight 54 provides the specificity of the meaning of a concept or term. The structural weight 55 assigns a degree of

significance to the concept or term. The corpus weight 56 inversely weighs the reference count, that is, the number of documents containing a concept or term at least once. Each score 52 is logarithmically compressed to provide a better linear vector representation and the scores are formed into normalized score vectors 57 for each of the documents 14.

The clustering module 43 forms clusters 58 of the documents 14 using the similarities of concepts and terms between the normalized score vectors 57, as further described below with reference to FIGURE 14. As a preparatory step in forming clusters 58, the clustering module 43 iteratively analyzes a set of seed candidate documents 60 to form a set of seed documents 59 from which the clusters 58 are generated.

The display and visualization module 44 complements the operations performed by the document analyzer 31 by presenting visual representations of the information extracted from the documents 14. The display and visualization module 44 generates a concept graph 61 of concept references determined over all documents 14, as further described below with reference to FIGURE 13.

Each module is a computer program, procedure or module written as source code in a conventional programming language, such as the C++ programming language, and is presented for execution by the CPU as object or byte code, as is known in the art. The various implementations of the source code and object and byte codes can be held on a computer-readable storage medium or embodied on a transmission medium in a carrier wave. The document analyzer 31 operates in accordance with a sequence of process steps, as further described below with reference to FIGURE 5.

Document Scoring

FIGURE 3 is a data flow diagram 65 showing the stages of document scoring performed by the document analyzer 14 of FIGURE 1. Document records 48 are preprocessed and noun phrases are extracted as concepts 65 and terms 66 for storage in concept records 49 and term records 50, respectively (transition 66). The concepts 65 and terms 66 are cataloged into content records 51 (transmission 67). A score 52 is then generated based on the frequencies 53, concept weights

54, structural weights 55, and corpus weights 56 of each concept 65 and term 66 (transitions 68-72). Optionally, a concept graph 61 can be generated (transition 73).

Document Clustering

FIGURE 4 is a data flow diagram showing the stages 75 of document clustering performed by the document analyzer 14 of FIGURE 1. Candidate seed documents 60 are selected to identify those documents 14 containing concepts 49 and, if necessary, terms 50, which represent categories of subject matter for potential clusters 52. The candidate seed documents 60 are evaluated (transition 76) based on similarity to a set of cluster centers 58, as measured by cosine values between normalized score vectors 57. Non-seed documents 78, that is, each of the documents 14 not selected as a seed document 60, are evaluated (transition 77) based on similarity to the set of cluster centers 58. Those non-seed documents 78 meeting a best fit criterion, subject to a minimum fit criterion, are formed (transition 79) into clusters 58.

Method Overview

FIGURE 5 is a flow diagram showing a method 80 for performing efficient document scoring and clustering, in accordance with the present invention. The method 80 is described as a sequence of process operations or steps, which can be executed, for instance, by a document analyzer 31 (shown in FIGURE 1).

As a preliminary step, the set of documents 14 to be analyzed is preprocessed (block 81) to identify terms and to extract concepts 65 and terms 66, as further described below with reference to FIGURE 6. Once preprocessed, the concepts 65 and terms 66 from the documents 14 are scored (block 82), as further described below with reference to FIGURE 11, and formed into clusters 58 (block 83), as further described below with reference to FIGURE 14. Optionally, the concept references can be displayed and visualized as a concept graph 61 (block 84), as further described below with reference to FIGURE 13. The routine then terminates.

Document Parsing

FIGURE 6 is a flow diagram showing the routine 90 for performing document parsing for use in the method 80 of FIGURE 5. The purpose of this routine is to retrieve a set of documents 14 from the storage device 13, identify terms occurring in each of the documents 14, and extract concepts 65 and terms 66 in the form of noun phrases for storage as concept records 49 and term records 50 in the database 30.

The set of documents 14 maintained in the storage device 13 is processed in an iterative processing loop (blocks 91-99). During each iteration (block 91), each document 14 is retrieved from the storage device 13 and converted into a document record 48 maintained in the database 30, as further described below with reference to FIGURE 7. The process of converting a document 14 into a document record 48 includes parsing through each document structure, that is, structural location, and creating a standardized representation of the document 14 to enable efficient, application-independent processing of the contents of each document 14.

Preliminarily, each document 14 may be preprocessed to remove extraneous formatting characters, such as hard returns or angle brackets, often used to embed previous email messages. Preprocessing maximizes syntactical extraction of desired terms and phrases without altering any semantic contents.

The global stop concept cache 45 contains a set of globally-applicable stop concepts used to suppress generic terms, such as "late," "more," "good," or any user-defined stop concepts, which are suppressed to emphasize other important concepts in specific review contexts. In the described embodiment, the global stop concept cache 45 is generated dynamically after document analysis as document review progresses. Other forms of term and concept exclusion could be provided, as would be recognized by one skilled in the art.

Next, terms within the documents 14 are identified (block 94). Terms are defined on the basis of extracted noun phrases, although individual nouns or tri-grams (word triples) could be used in lieu of noun phrases. In the described embodiment, the noun phrases are extracted using the LinguistX product licensed

by Inxight Software, Inc., Santa Clara, California. The identified phrases consist of regular nouns, as well as proper nouns or adjectives.

Next, the phrases are normalized (block 95) and used to identify canonical concepts (block 96). Unless indicated otherwise, the term "concepts" refers to canonical concepts as stored in a concept record 49 and applies equally to both concepts 65 and terms 66. Canonical concepts include the concepts 65 and terms 66 preferably processed into word stem form. In addition, the individual terms 66 comprising each concept 65 are converted to uniform lower case type and are alphabetized. By way of example, the sentence, "I went to the Schools of Business," would yield the canonical concept "business, school." Similarly, the sentence, "He went to Business School," would yield the same canonical concept "business, school." Other forms of canonical concepts could be used, including alternate word forms and arrangements, as would be recognized by one skilled in the art.

The canonical concepts are then used to build concept records 49 and term records 50 (block 97), as further described below with reference to FIGURE 8. Finally, content records 51 for each concept occurrence are built or updated (block 98), as further described below with reference to FIGURE 10. Processing continues with the next document 14 (block 99), after which the routine returns.

Document Record Schema

FIGURE 7 is a data structure diagram showing a schema 100 for a document record 101 maintained in the database 30 of FIGURE 1. One document record 101 is maintained per document 14. Each document record 101 uniquely identifies the associated document 14 and stores the contents of the message 14, preferably including any formatting and layout information, in a standardized representation. Each document record 101 includes fields for storing a document identifier (Doc ID) 102 and document name (Doc Name) 103.

Concept Record Schema

FIGURE 8 is a data structure diagram showing a schema 110 for a concept record 111 maintained in the database 30 of FIGURE 1. One concept record 111

is maintained per canonical concept. A canonical concept can include both concepts 65 and terms 66 arranged in alphabetized, normalized form. Each concept record 111 includes fields for storing a unique concept identifier (Concept ID) 112 and concept 113.

Associated Concept Record Schema

FIGURE 9 is a data structure diagram showing a schema 115 for an associated concept record 116 maintained in the database 30 of FIGURE 1. Concepts 65 consisting of more than one term 66 have associated records 116 stored as pairs of concept identifiers (Concept ID) 117 and term identifiers (Term ID) 118.

Content Record Schema

FIGURE 10 is a data structure diagram showing a schema 120 for a content record 121 maintained in the database 30 of FIGURE 1. One content record 121 is maintained per concept occurrence per structure per document 14. In addition, additional content records 121 can be maintained per additional concept occurrences per document 14. Thus, one document 14 could have an associated set of one or more content records 121 for each concept 65 or term 66 identified within the document 14. Similarly, one document 14 could have several associated sets of one or more content records 121 where, for instance, the concept 65 or term 66 appears in structurally distinct sections of the document 14, such as in the subject, title or body of a document. Each content record 121 includes fields for storing a document identifier (Doc ID) 122, concept identifier (Concept ID) 123, frequency 124, and structure code 125. The frequency 124 records the number of times that the concept 65 or term 66 is referenced within the document 14. The structure code 125 indicates the structural location within the document 14 from which the concept was extracted.

Document Scoring Routine

FIGURE 11 is a flow diagram showing a routine 130 for comparing documents 14 for use in the method 80 of FIGURE 5. The purpose of this routine

is to create a normalized score vector 57 for each document 14 and calculate a similarity metric between each of the normalized score vectors 57.

As an initial step, each concept 56 and term 66 is individually scored (block 131), as further described below with reference to FIGURE 12. Next, a normalized score vector 57 is created for each document 14 in an iterative processing loop (block 132-136). One document 14 is processed per iteration (block 132) by first creating the normalized score vector 57 (block 133). Each normalized score vector 57 includes a set of paired values, consisting of a concept identifier 112 for each concept 65 and term 66 occurring in that document 14 and the scores 52 for that concept 65 or term 66. Preferably, the paired values are ordered. In the described embodiment, only non-zero scores are maintained for efficiency.

For example, assume a normalized score vector 57 for a first document *A* is $\vec{S}_A = \{(5, 0.5), (120, 0.75)\}$ and a normalized score vector 57 for another document *B* is $\vec{S}_B = \{(3, 0.4), (5, 0.75), (47, 0.15)\}$. Document *A* has scores corresponding to concepts '5' and '120' and Document *B* has scores corresponding to concepts '3,' '5' and '120.' Thus, these documents only have concept '5' in common.

An inner product of the normalized score vector 57 for the current document 14 is calculated against the normalized score vectors 57 of each other document 14 among corresponding dimensions (block 134) by iterating through the paired values in the normalized score vector 57 to identify commonly occurring concepts 65 and terms 66. Cosine cos σ is equivalent to the inner products between two normalized vectors. The cosine cos σ provides a measure of relative similarity or dissimilarity between the concepts 65 and terms 66 occurring in each document 14 and can therefore serve as a form of similarity metric, as would be recognized by one skilled in the art. In the described embodiment, the cosine cos σ is calculated in accordance with the equation:

$$\cos\sigma_{AB} = \frac{\left\langle \vec{S}_A \cdot \vec{S}_B \right\rangle}{\left| \vec{S}_A \right\| \vec{S}_B \right|}$$

where $\cos\sigma_{AB}$ comprises the similarity for between the document A and the document B, $\vec{S}_A$ comprises a score vector 57 for document A, and $\vec{S}_B$ comprises a score vector 57 for document B. Other forms of determining a relative similarity metric are feasible, as would be recognized by one skilled in the art. Processing continues with the next document 14 (block 135), after which the routine returns.

Concept and Term Scoring Routine

FIGURE 12 is a flow diagram showing a routine 140 for scoring concepts 65 and terms 66 for use in the routine 130 of FIGURE 11. The purpose of this routine is to evaluate a score 52 for each concept 65 and term 66 based on frequency 53, concept weight 54, structural weight 55, and corpus weight 56. Each evaluated score 52 is compressed to enable better linear vector representation of those documents 14 which include lengthy contents.

A score 52 is calculated for each concept 65 and term 66 in an iterative processing loop (block 141-147). During each iteration (block 141), a score 52 is calculated as follows. First, a concept weight 54 is determined for the concept 65 or term 66 (block 142). The concept weight 54 reflects the specificity of the meaning of a single concept 65 or term 66.

In the described embodiment, each concept weight 54 is based on the number of individual terms 66 that make up the concept 65 or term 66. Each concept weight 54 is calculated in accordance with the following equation:

$$cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \le t_{ij} \le 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \le t_{ij} \le 6 \\ 0.25, & t_{ij} \ge 7 \end{cases}$$

where cw_{ij} comprises the concept weight and t_{ij} comprises a number of terms for occurrence j of each such concept i. The specificity of the meaning of a single concept 65 increases as the number of terms 66 occurring in the concept 65 increases. Intuitively, three to four terms 66 per concept 65 have proven more useful than other numbers of terms for differentiating between documents 14. Conversely, long concepts having in excess of five or more terms 66 tend to reflect parsing errors or are too specific for effective clustering.

Next, a structural weight 55 is determined for the concept 65 or term 66 (block 143). Each structural weight 55 reflects a varying degree of significance assigned to the concept 65 or term 66 based on structural location within the document 14. For example, subject lines in electronic mail (email) messages are assigned more importance than signature blocks.

In the described embodiment, each structural weight 55 is determined in accordance with the equation:

$$sw_{ij} = \begin{cases} 1.0, & if(j \approx SUBJECT) \\ 0.8, & if(j \approx HEADING) \\ 0.7, & if(j \approx SUMMARY) \\ 0.5 & if(j \approx BODY) \\ 0.1 & if(j \approx SIGNATURE) \end{cases}$$

where sw_{ij} comprises the structural weight for occurrence j of each such concept i. Other assignments of structural weight based on the location or arrangement of a concept 65 or term 66 occurrence within a document 14 are feasible, as would be recognized by one skilled in the art.

Next, a corpus weight is determined for the concept 65 or term 66 (block 144). The corpus weight 56 inversely weighs the reference count of the occurrences of each concept 65 or term 66 within a given document 14. The overall goal of forming clusters 58 is to group those documents 14 having similar content. Accordingly, the reference count of each concept 65 and term 66 can be used to differentiate document similarities. However, frequently referenced concepts 65 and terms 66 can dilute the differentiating measure of the reference counts and are ineffective in grouping similar documents. The reference counts for infrequently referenced concepts 65 and terms 66 also lack appreciable meaning as a differentiating measure except when evaluating clusters 58 for a small document set.

In the described embodiment, each corpus weight 56 is determined in accordance with the equation:

$$rw_{ij} = \begin{cases} \left(\frac{T - r_{ij}}{T}\right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

where rw_{ij} comprises the corpus weight, r_{ij} comprises a reference count for occurrence j of each such concept i, T comprises a total number of reference counts of documents in the document set, and M comprises a maximum reference count of documents in the document set. A value of 10% is used to indicate the maximum reference count at which a score contribution is discounted, although other limits could be used, as would be recognized by one skilled in the art.

Next, the actual score 52 for each concept 65 and term 66 is determined (block 145). Note each concept 65 and term 66 could occur one or more times within the same document 14 and could be assigned different structural weights 55 based on structural locations within the document 14. Each score 52 represents the relative weight afforded to each concept 65 and term 66 with respect to a particular document 14.

In the described embodiment, each score 52 is calculated in accordance with the equation:

$$S_i = \sum_{1 \to n}^{j} f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

where S_i comprises the score 52, f_{ij} comprises the frequency 53, $0 < cw_{ij} \leq 1$ comprises the concept weight 54, $0 < sw_{ij} \leq 1$ comprises the structural weight 55, and $0 < rw_{ij} \leq 1$ comprises the corpus weight 56 for occurrence j of concept i within a given document 14. Finally, the score 52 is compressed (block 146) to minimize the skewing caused by concepts 65 and terms 66 occurring too frequently.

In the described embodiment, each compressed score is determined in accordance with the equation:

$$S_i' = \log(S_i + 1)$$

where S_i' comprises the compressed score 52 for each such concept i. Logarithmical compression provides effective linear vector representation of

those documents 14 having a large body of content. Other forms of score compression could be used, as would be recognized by one skilled in the art.

Processing continues with the next concept 65 or term 66 (block 147), after which the routine returns.

Concept Reference Frequencies Graph

FIGURE 13 is a graph 150 showing, by way of example, the frequency of concept references. The graph 150 illustrates the effect of inversely weighing the reference counts of concepts 65 and terms 66. The *x*-axis represents the individual concepts 65 and terms 66 occurring in the set of documents 14. The *y*-axis indicates the reference counts 152, that is, the number of documents 14 containing a given concept 65 or term 66. A curve 155 reflects the ratio of concepts and terms versus reference counts. Accordingly, the concepts 65 and terms 66 appearing in at least 10% of the documents are discounted as lacking sufficient differentiating characteristics. A line 156 reflects the 10% cutoff point and the curve 153 reflects the corpus weight 56 of each of the concepts 65 and terms 66, up to the 10% cutoff point 154.

Cluster Forming Routine

FIGURE 14 is a flow diagram showing a routine 160 for forming clusters 58 for use in the method 80 of FIGURE 5. The purpose of this routine is to use the scores 52 of the concepts 65 and terms 66 as stored into the normalized score vectors 57 to form clusters 58 of documents 14 based on relative similarity.

The routine proceeds in two phases. During the first phase (blocks 161-169), seed candidate documents 60 are evaluated to identify a set of seed documents 59. During the second phase (blocks 170-176), non-seed documents 78 are evaluated and grouped into clusters 58 based on a best-fit criterion.

First, candidate seed documents 60 are identified (block 161) and ordered by category (block 162). In the described embodiment, the candidate seed documents 60 are selected based on a subjective evaluation of the documents 14 and are assigned into generalized categories, such as "responsive," "non-

responsive," or "privileged." Other forms of classification and categorization are feasible, as would be recognized by one skilled in the art.

Next, the candidate seed documents 60 are ordered within each category based on length (block 163). Each candidate seed document 60 is then processed in an iterative processing loop (blocks 164-169) as follows. The similarity between each current candidate seed document 60 and the cluster centers 58, based on seed documents already selected 59, is determined (block 165) as the cosine cos σ of the normalized score vectors 57 for the candidate seed documents 60 being compared. Only those candidate seed documents 60 that are sufficiently distinct from all cluster centers 58 (block 166) are selected as seed documents 59 (block 167). In the described embodiment, a range of 0.10 to 0.25 is used, although other ranges and spatial values could be used, as would be recognized by one skilled in the art.

If the candidate seed documents 60 being compared are not sufficiently distinct (block 166), the candidate seed document 60 is grouped into a cluster 58 with the most similar cluster center 58 to which the candidate seed document 60 was compared (block 168). Processing continues with the next candidate seed document 60 (block 169).

In the second phase, each non-seed document 78 is iteratively processed in an iterative processing loop (blocks 170-176) as follows. The non-seed documents 78 are simply those documents 14 other than the seed documents 60. Again, the similarity between each current non-seed document 78 and each of the cluster centers based on the seed documents 59 is determined (block 171) as the cosine cos σ of the normalized score vectors 57 for each of the non-seed documents 78. A best fit between the current non-seed document 78 and the cluster centers 58 is found subject to a minimum fit criterion (block 172). In the described embodiment, a minimum fit criterion of 0.25 is used, although other minimum fit criteria could be used, as would be recognized by one skilled in the art. If a best fit is found (block 173), the current non-seed document 78 is grouped into the cluster 58 having the best fit (block 175). Otherwise, the current non-seed document 78 is grouped into a miscellaneous cluster (block 174).

Processing continues with the next non-seed document 78 (block 176). Finally, a dynamic threshold is applied to each cluster 58 (block 177), as further described below with reference to FIGURE 15. The routine then returns.

FIGURE 15 is a flow diagram showing a routine 180 for applying a dynamic threshold for use in the routine 160 of FIGURE 5. The purpose of this routine is to perform "tail cutting" to each cluster 58 by dynamically evaluating and strengthen membership on a cluster-by-cluster basis for use in a further embodiment. Tail cutting creates tighter clusters 58 by identifying and relocating "outlier" documents.

FIGURE 16 is a graph diagram 200 showing, by way of example, a dynamic threshold 204 in a cluster 201 of documents 202. The dynamic threshold 204 is based on an analysis of the similarities of the documents 202 from the center 203 of the cluster 201. Those documents 202 falling outside of the dynamic threshold 204, that is, outlier documents 205, are identified and relocated, if possible, to other clusters.

Referring back to FIGURE 15, in applying a dynamic threshold 204, each of the documents 202 in each of the clusters 201 is processed in a pair of iterative processing loops (blocks 181-184) as follows. During each iteration of the outer processing loop (block 181), a current cluster 201 is selected and, during each iteration of the inner processing loop (block 182), a document 202 is selected from the current cluster 201. The similarity to the center 203 of the current cluster 201 for each document 202 is calculated (block 183) and processing continues with the next document 202 (block 184).

Upon completion of the computation of similarity calculations for each document 202, the standard deviation of all documents 202 from the center 203 of the current cluster 201 is determined and a dynamic threshold 204 is set (block 185). In the described embodiment, a dynamic threshold 204 of $\pm$ 1.2 standard deviations is used, although other dynamic thresholds 204 could also be used, as would be recognized by one skilled in the art. Next, those documents 202 in the current cluster 201, which are outside of the dynamic threshold 204, that is, outlier documents 205, are identified (block 186) and are processed in an iterative

processing loop (blocks 187-193) as follows. The similarity between each outlier document 205 and each of the cluster centers is determined (block 188) based on the cosine cos σ of the normalized score vectors 57 for each of the outlier documents 205. A best fit between the outlier document 205 and the cluster centers is found subject to a minimum fit criterion and the dynamic threshold 204 (block 189). In the described embodiment, a minimum fit criterion of 0.25 is used, although other minimum fit criteria could be used, as would be recognized by one skilled in the art. The dynamic threshold 204 used to rescale each cluster-to-document similarity, which enables comparisons of similarities across all available clusters, is calculated in accordance with the equation:

$$similarity_{new} = \frac{similarity_{old}}{(1 - threshold)}$$

where $similarity_{new}$ comprises a new similarity, $similarity_{old}$ comprises the old similarity and *threshold* comprises the dynamic threshold 204.

If a best fit is found (block 190), the outlier document 205 is grouped into the cluster 58. Otherwise, the outlier document 205 is grouped into a miscellaneous cluster (block 191). Processing continues with the next outlier document 205 (block 192) and the next cluster 201 (block 193), after which the routine returns.

While the invention has been particularly shown and described as referenced to the embodiments thereof, those skilled in the art will understand that the foregoing and other changes in form and detail may be made therein without departing from the spirit and scope of the invention.